Online Learning

Ref: Chapter 8 of FOML book

Supervised learning: Training phase followed by Test phase

Online learning: No such separation

Supervised/PAC learning: Distributional assumptions

Online learning: No such assumptions are made

Online learning: Interaction

"Sequential", horizon: "T" rounds

for $t = 1, \dots T$

{

Algorithm receives $x_t \in \mathcal{X}$

& predicts $\hat{y}_t \in \mathcal{Y}$

Observe y_t & incur loss $L(\hat{y}_t, y_t)$

}

Goal of an online learning algorithm:

minimize cumulative loss $\sum_{t=1}^{T} L(\hat{y}_t, y_t)$

Examples:

① Classification problem: $\mathcal{Y} = \{0,1\}$

$L(y,y') = |y-y'|$

or $L(y,y') = \mathbb{1}\{y \neq y'\}$

② Regression

$L(y,y') = (y-y')^2$ with $\mathcal{Y} \subseteq \mathbb{R}$

"Prediction with expert advice"

In each round "t", algorithm $\mathcal{A}$ receives

$x_t \in \mathcal{X}$ & advice $y_{t,i} \in \mathcal{Y}$, $i=1,\dots,N$

N = # of experts

alg $\mathcal{A}$ predicts $\hat{y}_t$ & observes $L(\hat{y}_t, y_t)$

Goal: Minimize cumulative loss.

Practical motivation:-

Movie recommendation

Experts: IMDB, RT, your friend

For a movie, each expert makes a prediction "good" or "bad"

Notion of "regret":

(Cumulative) Loss of expert i: $\sum_{t=1}^{T} L(y_{t,i}, y_t)$

—"— best expert: $\min_{i=1\sim N} \sum_{t=1}^{T} L(y_{t,i}, y_t)$

—"— of alg A: $\sum_{t=1}^{T} L(\hat{y}_t, y_t)$

Regret $R_T = \sum_{t=1}^{T} L(\hat{y}_t, y_t) - \min_{i=1\sim N} \sum_{t=1}^{T} L(y_{t,i}, y_t)$

Realizable Case or the Case with the "perfect" expert:

Loss: 0-1 loss Binary classification

Want an algorithm with the smallest # of mistakes

To judge an algorithm A, we use the max # of mistakes done before zeroing in on the perfect expert

$$M_A(c) = \max_{x_1 -- x_T} |\text{mistakes}(A, c)|,$$

true concept

for some T after which there are no mistakes.

For a concept class $\mathcal{C}$,

$$M_A(\mathcal{C}) = \max_{c \in \mathcal{C}} M_A(c)$$

Halving algorithm

Have a set of active experts

Alg A prediction: majority vote

In case of mistake: remove all experts who made a wrong prediction from the active set.

Pseudo code: Active set $H_1 \in H$ (all experts)

for $t = 1 -- T$

{ Receive x_t

$\hat{y}_t \in \text{majority vote}(H_t, x_t)$

Observe y_t

If $y_t \neq \hat{y}_t$
{
$H_{t+1} \leftarrow \{c \in H_t : c(x_t) = y_t\}$
}
else
$H_{t+1} \leftarrow H_t$
}

<u>Mistake bound:</u>

$$M_{\text{Halving}}(H) \leq \log_2 |H|$$

<u>Proof:</u> On a mistake, the active set is reduced by atleast half.

Hence, after $\log_2 |H|$ mistakes, only one element remain in the active set & this can only be the perfect expert. (= target concept)

Non-realizable case & the weighted majority (WM) algorithm:

Suppose there is no perfect expert (or the true concept is not in the hypothesis set)

WM: maintain a weight for each expert

On a mistake, an expert with erroneous prediction gets his weight reduced.

WM prediction: majority vote but using weights

<u>Pseudocode:</u>

for $i = 1 \ldots N$ $\{ w_{1,i} = 1 \}$

for $t = 1 \ldots T$

$\{$ Receive x_t

weighted majority

If $\sum_{i : y_{t,i} = 1} w_{t,i} \geq \sum_{i : y_{t,i} = 0} w_{t,i}$, then

$\{$ predict $\hat{y}_t = 1 \}$

else $\{$ predict $\hat{y}_t = 0 \}$

Receive y_t

If $(\hat{y}_t \neq y_t)$, then

{

for $i = 1 \ldots N$

{

If $(y_{t,i} \neq y_t)$ $\{ w_{t+1,i} = \beta w_{t,i} \}$ $\beta \in [0,1)$

else $\{ w_{t+1,i} = w_{t,i} \}$

}

}

Mistake bound for WM algorithm:-

Let M_T be # of mistakes of WM algorithm

m_T^* ——"—— best expert

Let $W_t = \sum_{i=1}^{N} w_{t,i}$

If WM makes a mistake in round t, then

$$W_{t+1} \leq \left[\frac{1}{2} + \frac{\beta}{2} \right] W_t = \left(\frac{1+\beta}{2} \right) W_t$$

$W_1 = N$ and there are M_T # of mistakes after

T rounds, implying

$$W_T \le \left(\frac{1+\beta}{2}\right)^{m_T} W_1 = \left(\frac{1+\beta}{2}\right)^{m_T} N \quad (*)$$

To lower bound W_T:

For an expert i: $w_{T,i} = \beta^{m_{T,i}}$, where

$m_{T,i}$ = # of mistakes of expert i

$$W_T \geq w_{T,i} \text{ for any } i$$

In particular, $W_T \geq \beta^{m_T^*}$ → weight of best expert who made m_T^* mistakes after T rounds.

Combining the lower bound above with the bound in $(*)$, we obtain

$$\beta^{m_T^*} \le W_T \le \left(\frac{1+\beta}{2}\right)^{m_T} N$$

$$\Rightarrow m_T^* \log \beta \le \log N + m_T \log\left(\frac{1+\beta}{2}\right)$$

$$\Rightarrow m_T \log\left(\frac{2}{1+\beta}\right) \le \log N + m_T^* \log\left(\frac{1}{\beta}\right)$$

Hence,

$$m_T \leq \frac{\log N + m_T^* \log\left(\frac{1}{\beta}\right)}{\log\left(\frac{2}{1+\beta}\right)}$$

mistake bound for WM algorithm

$$m_T \leq O(\log N) + \text{const.} \times (\text{mistakes of best expert})$$

Note: No assumptions made on how the data is generated & the sequence of data points $\{(x_t, y_t)\}_{t=1 \ldots t}$ could be chosen in an adversarial fashion.

The Case of convex losses

loss function L bounded in $[0, 1]$ and is convex in the first argument.

$$L(\hat{y}, y)$$

parameter in which L is convex

$$\text{Regret} \quad R_T = \sum_{t=1}^{T} L(\hat{y}_t, y_t) - \min_{i=1 \ldots N} \sum_{t=1}^{T} L(y_{t,i}, y_t)$$

Recall: $N \to$ # of experts

$y_{t,i} \to$ prediction of expert i in round t

$\hat{y}_t \to$ prediction of the algorithm

"Exponential weighted average" (EWA)

Pseudocode:-

(1) Initialization

for $i = 1 \ldots N$

{ $w_{1,i} = 1$; }

(2) for $t = 1, \ldots T$

{

Receive x_t

Predict $\hat{y}_t = \dfrac{\sum_{i=1}^{N} w_{t,i}\, y_{t,i}}{\sum_{i=1}^{N} w_{t,i}}$

Receive y_t & incur loss $L(\hat{y}_t, y_t)$

//update weights: for $i = 1 \ldots N$

{

Parameter to be specified later

$$w_{t+1,i} = w_{t,i} \exp(-\eta \underbrace{L(y_{t,i}, y_t)}_{\text{loss of expert } i})$$

}

}

Note:

$$w_{t+1,i} = w_{t,i} \exp\left(-\eta L(y_{t,i}, y_t)\right)$$

$$= \exp\left(-\eta \sum_{s=1}^{t} L(y_{s,i}, y_s)\right)$$

Cumulative loss of expert i up to time t

$$= \exp\left(-\eta L_{t,i}\right)$$

The weight of expert i depends on the cumulative loss & the individual losses need not be stored by the EWA algorithm

Regret analysis for EWA algorithm:-

Theorem: L is convex & bounded in $[0,1]$

For any $\eta > 0$, and any sequence $y_1, \dots, y_T$, the regret of EWA after T rounds satisfies

$$R_T \le \frac{\log N}{\eta} + \frac{\eta T}{8}$$

Choosing $\eta = \sqrt{\dfrac{8 \log N}{T}}$, we obtain

$$R_T \leq \sqrt{\frac{T}{2} \log N}$$

Since $R_T \leq \sqrt{\frac{T}{2} \log N}$, $\dfrac{R_T}{T} \leq \dfrac{\sqrt{\log N}}{\sqrt{T} \times \sqrt{2}}$

$\dfrac{R_T}{T} \to 0$ as $T \to \infty$

(intuitively, the algorithm matches the prediction of the best expert at a rate

$\dfrac{R_T}{T} = \tilde{O}\left(\dfrac{1}{\sqrt{T}}\right)$ ignoring log factors

<u>Proof:</u>

Define $\Phi_t = \log \sum_{i=1}^{N} w_{t,i}$

Let P_t denote the distribution over $\{1, \dots N\}$

with $P_{t,i} = \dfrac{w_{t,i}}{\sum_{i=1}^{N} w_{t,i}}$

$$\Phi_{t+1} - \Phi_t = \log\left(\frac{\sum_{i=1}^{N} w_{t+1,i}}{\sum_{i=1}^{N} w_{t,i}}\right)$$

$$= \log\left[\sum_{i=1}^{N} \left(\frac{w_{t,i}}{\sum_{i=1}^{N} w_{t,i}} \right) \exp\left(-\eta L(y_{t,i}, y_t)\right) \right]$$

$$= \log\left[\sum_{i=1}^{N} P_{t,i} \exp\left(-\eta L(y_{t,i}, y_t)\right) \right]$$

$$= \log\left(E_{P_t}\left(\exp(\eta X) \right) \right),$$

where $X = -L(y_{t,i}, y_t)$

Recall: Hoeffding's Lemma.

For a r.v. X with mean μ, $a \le X \le b$

$$E\left(\exp(t(X-\mu))\right) \le \exp\left(\frac{t^2 (b-a)^2}{8} \right)$$

Jensen's inequality:

For a r.v. X with mean $E(X) < \infty$, and convex function f, we have

$$f(E(X)) \le E(f(X))$$ e.g. $EX^2 \ge (EX)^2$

Back to EWA regret bound!

$$\Phi_{t+1} - \Phi_t = \log\left(\mathbb{E}_{p_t}(\exp(\eta x))\right)$$

$$= \log \mathbb{E}_{p_t}\left(\exp\left(\eta(x - \mathbb{E}x) + \eta \mathbb{E}x\right)\right)$$

apply Hoeffding's lemma to this part (note: $-L \in [-1, 0]$)

$$\leq \log \exp\left(\frac{\eta^2}{8} + \eta \mathbb{E}_{p_t} x\right)$$

$$= \frac{\eta^2}{8} + \eta \mathbb{E}_{p_t} x$$

$$= \frac{\eta^2}{8} - \eta \mathbb{E}_{p_t}\left(L(y_{t,i}, y_t)\right)$$

Jensen's inequality

$$\leq \frac{\eta^2}{8} - \eta L\left(\mathbb{E}_{p_t}(y_{t,i}), y_t\right)$$

$$= \frac{\eta^2}{8} - \eta L(\hat{y}_t, y_t)$$

So, we have

$$\Phi_{t+1} - \Phi_t \leq \frac{\eta^2}{8} - \eta L(\hat{y}_t, y_t)$$

$$\Phi_{T+1} - \Phi_1 \leq \frac{\eta^2 T}{8} - \eta \sum_{t=1}^{T} L(\hat{y}_t, y_t)$$

So far, we have derived an upper bound on $\Phi_{T+1} - \Phi_1$

We will lower bound $\Phi_{T+1} - \Phi_1$.

$\Phi_{T+1} - \Phi_1$ total weight of expert i upto time t

$$= \log \sum_{i=1}^{N} \exp(-\eta L_{T,i}) - \log N$$

sum > max for +ve numbers

$$\geq \log \max_{i=1..N} \exp(-\eta L_{T,i}) - \log N$$

$$= -\eta \min_{i=1..N} L_{T,i} - \log N$$

relate $\log \max(e^{-a}, e^{-b})$, to $-\min(a,b)$, $a>0$, $b>0$

Combining upper & lower bounds

$$-\eta \min_{i=1..N} L_{T,i} - \log N \leq \Phi_{T+1} - \Phi_1 \leq \frac{\eta^2 T}{8} - \eta \sum_{t=1}^{T} L(\hat{y}_t, y_t)$$

Re-arranging,

$$\sum_{t=1}^{T} L(\hat{y}_t, y_t) - \min_{i=1..N} L_{T,i} \leq \frac{\log N}{\eta} + \frac{\eta T}{8}$$

Hence proved. ∎

Perceptron algorithm

Assumption: $S = \{(x_i, y_i),\ i = 1 \dots T\}$ is linearly separable

Pseudo-code

Initialization: $w_1 \leftarrow w_0$ // usually $w_0 = 0$

for $t = 1 \dots T$

{

Receive x_t

Predict $\hat{y}_t = \mathrm{sgn}(w_t^T x_t)$

Observe y_t

If $(\hat{y}_t \neq y_t)$

{

$w_{t+1} = w_t + y_t x_t$

}

else $w_{t+1} = w_t$

}

Return w_{T+1}

Recall: Data "S" is linearly separable

$$\exists w^* \quad \text{s.t.} \quad x_i^T w^* > 0 \text{ if } y_i = 1 \\ < 0 \text{ if } y_i = -1 \quad \Big\} \, i = 1 \ldots T$$

$$w_{t+1} = w_t + \Delta w_t$$

x_t is correctly classified

$$\Delta w_t = \begin{cases} 0 & \text{if } w_t^T x_t > 0 \ \& \ y_t = 1 \ (\text{or}) \\ & \quad w_t^T x_t < 0 \ \& \ y_t = -1 \\ x_t & \text{if } w_t^T x_t \le 0 \ \& \ y_t = 1 \\ -x_t & \text{if } w_t^T x_t \geq 0 \ \& \ y_t = -1 \end{cases}$$

$\rightarrow$ error misclassification

To understand the update rule, observe that

when $w_t^T x_t \le 0 \ \& \ y_t = 1$,

$$w_{t+1}^T x_t = w_t^T x_t + x_t^T x_t$$

$$= w_t^T x_t + \|x_t\|^2 \geq w_t^T x_t$$

In the other error case, we have

$$w_{t+1}^T x_t \le w_t^T x_t \quad \text{when } w_t^T x_t \geq 0 \ \& \ y_t = -1$$

An objective function minimized by Perceptron is

$$F(w) = \frac{1}{T} \sum_{t=1}^{T} \max(0, -y_t(w^T x_t))$$

convex function

since (i) $w \to -y_t(w^T x_t)$ is convex

(ii) max is convex

Perceptron can be seen as a subgradient descent for the objective F.

Perceptron convergence

(A1) Let $x_1, \dots x_T \in \mathbb{R}^d$ be a sequence of points with $\|x_t\| \leq r \;\; \forall t = 1 \dots T$.

(A2) Suppose $\exists\, \rho > 0, \; w^* \in \mathbb{R}^d$ s.t.

$$\rho \leq \frac{y_t(w^{*T} x_t)}{\|w^*\|}, \quad t = 1, \dots T$$

margin

Then, the number of updates of Perceptron, when seeing $x_1, \ldots x_T$ is bounded by $\frac{\gamma^2}{\rho^2}$

Pf: Let J be the subset of T rounds at which there is an update & let M be the total # of updates, i.e., $|J| = M$.

$$M\rho \leq \frac{w^{*T}\left(\sum_{t\in J} y_t x_t\right)}{\|w^*\|}$$

Cauchy-Schwarz

$$\leq \left\| \sum_{t\in J} y_t x_t \right\|$$

$$= \left\| \sum_{t\in J} (w_{t+1} - w_t) \right\|$$

telescopic sum

$$= \| w_{T+1} \| \quad (\text{assuming } w_0 = 0)$$

$$= \sqrt{\sum_{t\in J} \left(\|w_{t+1}\|^2 - \|w_t\|^2\right)}$$

telescopic sum

$$= \sqrt{\sum_{t\in J} \|w_t + y_t x_t\| - \|w_t\|^2}$$

$$= \sqrt{\sum_{t \in J} 2 y_t w_t^T x_t + \|x_t\|^2}$$

$$\leq \sqrt{\sum_{t \in J} \|x_t\|^2} \quad \text{since } y_t w_t^T x_t \leq 0$$

$$M e \leq \sqrt{M r^2} \quad \text{by (A1) \& } |J| = M$$

So, $\sqrt{M} \leq \frac{r}{e}$ or

$$\boxed{M \leq \frac{r^2}{e^2}}$$

Remark :-

$$W_{T+1} = \sum_{t \in J} y_t x_t \quad \text{since } w_{t+1} = w_t + y_t x_t \text{ for } t \text{ in } J$$

W_{T+1} is a linear combination of a subset of vectors in the dataset

& these vectors $\{x_t \mid t \in J\}$ can be seen as the equivalent of support vectors for perceptron.

Reading assignment

Kernel perceptron. (figure 8.9 of FOML book)

Predict: $\hat{y}_t = \mathrm{sgn}\left(\sum_{s=1}^{T} \alpha_s y_s K(x_s, x_t)\right)$

Check the update rule for α_t